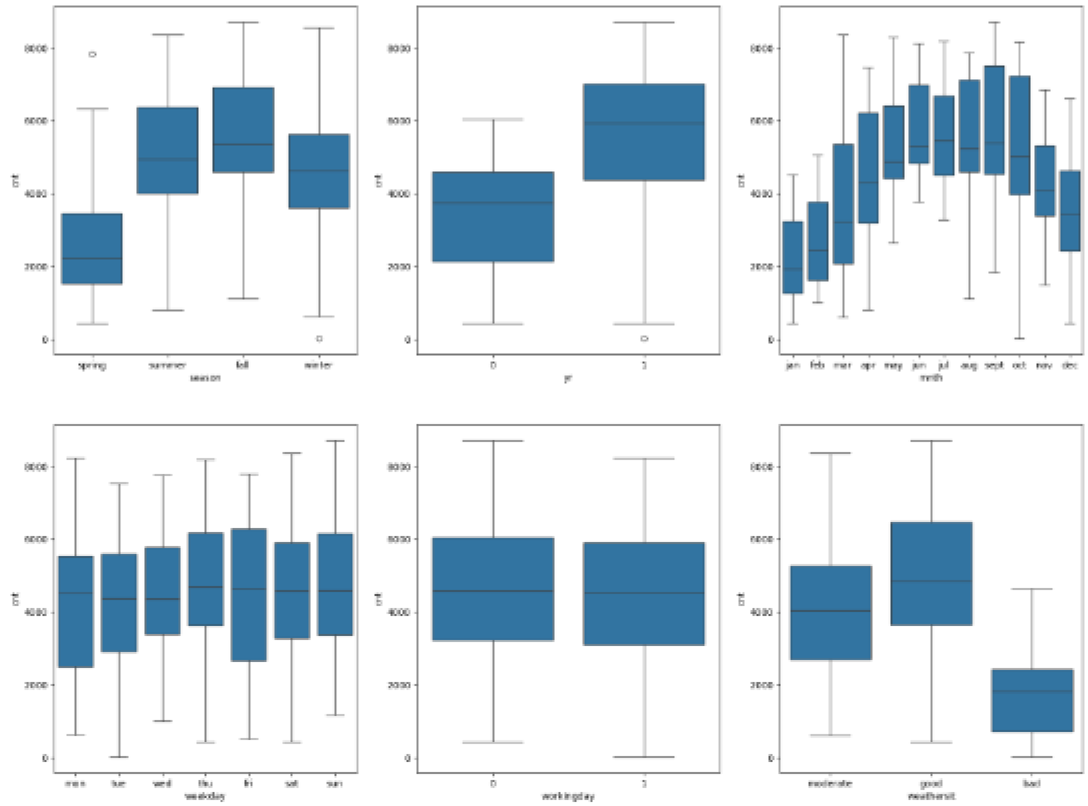
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans: Season, month, weekday, year, weathersit and working day are the categorical variables available in the dataset. These variables have a great impact on the dependent variable ‘cnt’.



1. Demand for rental bikes is highest during fall Season

2. Demand has grown in 2019

3. We can see growing demand from March till June and after a drop there is a highest demand in the month of September. Demand has been decreasing after September month of every year.

4. when weathersit is clear,we can see highest demand.

5. Comparatively Thursday and Friday, i.e. towards the weekend demand increases.

2. Why is it important to use **drop_first=True** during dummy variable creation?

Ans:drop_first = True, reduces the extra column created during dummy variable creation by finding the correlation among dummy variables. Thus, an unwanted or irrelevant column creation is reduced hence more relevant and readable data set is created.

It indicates if k−1 dummy variables have to be created from k categorical levels by omitting the first level.

Consider the example of married, in relationship, single. Here there are 3 types of values for categorical column status. If a person is not married and not in relationship, he/she will be single similarly if the person is neither single and nor married then he/she will be in a relationship. So, there is no need of all 3 variables. Hence we only need k-1 variables while creating a dummy variable.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans: The variables Atemp and Temp have the highest correlation with the target variable among the numerical variables.

4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans: After building the model on the training set, the validationof assumption of linear regression was done through the below findings:

- Normality of error (normal distributed of error terms).
- Visibility of linearity among variables
- In residual values, no visible patterns were seen
- Auto-correlation was not found
- Considering multi-collinearity among variables are insignificant

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans: The top 3 features contributing significantly towards explaining the demand of the shared bikes are

- Weathersit_bad (negative correlation)
- Temperature(Season)
- Year

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans: It is a statistical method which is used to model the relationship between a dependent variable and one or more independent variables. The Objective of this algorithm is to predict the probable outcome of the dependent variable based on the values of the independent variables.

a.Components in a Linear regression:

1.Dependent Variable (Target): The variable we want to predict (e.g., sales, price, Durability).

2.Independent Variables (Features): The variables used for prediction (e.g., advertising spend, number of employees, usage time).

Linear Equation: The relationship is modeled as:$Y=\beta0+\beta1X1+\beta2X2+\ldots+\beta nXn+\epsilon$

where:

- Y = predicted value
- $\beta0$ = y-intercept
- $\beta1,\beta2,\ldots,\beta n$ = coefficients for each feature
- $X1,X2,\ldots,XnX$ = independent variables
- ϵ = error term (residuals)

b. Assumptions of Linear Regression

To ensure valid results, we assumethat:

- Linearity: The relationship between independent and dependent variables are linear.
- Independence: The Observations are independent to each other.
- Homoscedasticity: The variance of the residuals remains constant across all values of X
- Normality: The residuals are normally distributed.

c. Steps in the Linear Regression Algorithm

1. Data Collection: Gatheringof data that are relevant to the dependent and independent variables.

2. Data Preprocessing:
   - Handling Missing Values: Remove or replace missing data.
   - Feature Scaling: Scale features whenever required (Not mandatory for linear regression).
   - Categorical Variables: Conversion of categorical variables into useful dummy/indicator variables if needed.

3. Model Training:

   Calculating Coefficients using methods like Ordinary Least Squares (OLS), which minimizes the Total amount of squared differences between the observed and predicted values

4. Model Evaluation:

   Use metrics such as:
   - R-squared
   - Mean Squared Error (MSE).

- Root Mean Squared Error (RMSE)
- Adjusted R-squared

d. Model Prediction: Once the model is trained and evaluated, use it to make predictions on new data.

e. Residual Analysis: Examination of residuals to validate assumptions and check for any patterns that might suggest a violation of assumptions.

2. Explain the Anscombe’s quartet in detail.

Ans. It is a group of four datasets that have almost identical statistical properties like same mean, variance, and correlation but they will differ widely in their distribution and graphical representation. It was created to illustrate the importance of graphical analysis in data analysis and effect of other outliners by the statistician Francis Anscombe in 1973.

3. What is Pearson’s R?

Ans. Pearson's Correlation coefficient, often referred to as Pearson's R, quantifies the strength of the correlation between two variables and is frequently employed in linear regression. It is the ratio between the covariance of two variables and the product of their standard deviations.

Itranges from -1 to +1, with +1 signifying a perfect positive linear correlation and -1 indicating a perfect negative linear correlation. Values between these extremes defines the degree of collinearity between the two variables.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans.Scaling is priorly used for ensuring that a model functions effectively with the right range of coefficients. For instance, if you have two independent variables namely price and months, on which sale of an automobile depend on, the price could vary significantly while months only range from 1 to 12. By Properly scaling the price variable we can prevent decimal errors in the model.
There are two major types of scaling:

1. **Normalized Scaling**: This type of scaling converts the data distribution into a Gaussian shape without a fixed range. It is commonly used in neural networks.
2. **Standardized Scaling**: This is exemplified in the previous scenario, where the values of the variables are compressed into a specific range that aligns with the model’s requirements.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans. This scenario happens when the correlation is perfect . The value of VIF will be large when there is a correlation between variables. If the value of VIF comes out to be 4, it means that since there is a multi-collinearity present the variance of model coefficient is inflated by factor of 4.
For perfect correlation, r2 is 1 which leads to 1/1-1 = infinity . To solve such cases, we drop one of the variables which is causing perfect collinearity hence eliminate the case of VIF being infinite.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression

Ans.**Q-Q (Quantile-Quantile) plot** is a graphical tool used to assess whether a dataset follows a particular distribution, mostly the normal distribution. In linear regression, a Q-Q plot is used to validate the assumption of normality of the residuals.

Use and Importance:

1. **Normality Check**: The Q-Q plot compares the quantiles of the residuals against the quantiles of a normal distribution. If the residuals follow a normal distribution, the points will approximately lie on the reference line.
2. **Identifying Deviations**: Deviations from the line in the Q-Q plot can indicate non-normality in the residuals. This can manifest as heavy tails (indicating outliers) or skewness.
3. **Model Validation**: Checking the normality of residuals helps validate the assumptions of linear regression, ensuring the model's reliability for inference.
4. **Transformations**: If the Q-Q plot shows significant deviations from normality, it may indicate the need for data transformation (e.g., log, square root) to meet the linear regression assumptions.